10.1515/astro-2017-0129

Baltic Astronomy, vol. 25, 267–274, 2016

KERNEL DENSITY ESTIMATION IN THE STUDY OF STAR CLUSTERS

Anton F. Seleznev
Astronomical Observatory, Ural Federal University,
Mira str. 19, Ekaterinburg, 620002, Russia; anton.seleznev@urfu.ru

Received: 2016 September 27; accepted: 2016 October 17

Abstract. The kernel estimator method is used to evaluate the surface and spatial star number density in star clusters. Both density maps and radial density profiles are plotted. These estimates are used to derive the cluster size, the number of cluster stars and the cluster mass, and to study the cluster structure. The kernel estimator is also used to plot the luminosity function, mass function, the velocity distribution, and Hess diagrams for star clusters. The advantages of the kernel estimator method and technical details of its use are illustrated by modern results for the open cluster NGC 4337.

Key words: methods: statistical – globular clusters: general – open clusters and associations: general

1. INTRODUCTION

Astronomers often deal with functions, which either represent the probability density or are closely related to it. These include surface and spatial star number density, phase-space density and velocity distribution function, luminosity function, mass function, etc. Very often the only difference between a function and a probability density is normalization. The probability density is normalized to unity, and the luminosity function, for example, is normalized to the number of stars in the field studied. The relation of radial profiles of the surface and spatial density to probability density functions is somewhat more complicated (Seleznev 2016).

Hence, methods for evaluating probability density can be used to determine other functions of astronomical interest. Such methods were described in Silverman (1986). They include histograms, naive estimator, kernel estimator and variable (adaptive) kernel method, the nearest neighbor method, orthogonal series estimators, maximum penalized likelihood estimators, and general weight function estimators.

Of course, histogram is the most popular method. It is indeed very simple and visually clear. However, it has a number of disadvantages. Moreover, there is some sort of misunderstanding concerning a histogram, which distorts its use. First, the final estimate is not a continuous and differentiable function. Second, the appearance of a histogram depends not only on the bin width, but also on the initial point selection. Third, it is necessary to take into account the interval error associated with the histogram (Kholopov 1953). The misunderstanding mentioned

above concerns deriving the standard deviation of the histogram bin values. The common practice is to use the square root of N (where N is the number of stars in the bin), or the square root of N divided by the ring area (if stars are counted in the concentric rings for deriving the radial surface density profile) as a standard deviation of the histogram bin value. It is usually substantiated by considerations that star counts obey Poisson statistics. However, when a star cluster is studied, N is usually the sum of the number of cluster stars and the number of field stars. One can assume that the number of field stars obeys Poisson statistics, however, the number of cluster stars obviously obeys a different distribution law. The correct method for evaluating the standard deviation was proposed by Vasilevsky (1985). On the other hand, star counts in the fields of ten open star clusters (Danilov, Matkin & Pylskaya 1985) revealed that the standard deviation of the number of stars depends on the mean number of field stars as a power law with an exponent of $\alpha = 0.89 \pm 0.04$ because of large-scale fluctuations of the background density.

The kernel estimator method, which is intuitive and easy to implement, is free from disadvantages of a histogram. The kernel estimator was used to evaluate the luminosity function (Seleznev 1998; Seleznev et al. 2000; Prisinzano et al. 2001), to derive and analyze surface density maps in star clusters (Kirsanova et al. 2008; Seleznev et al. 2010; Carraro et al. 2012), to derive surface and spatial radial density profiles of star clusters and their numerical models (Merritt & Tremblay 1994; Danilov, Putkov & Seleznev 2014; Carraro et al. 2016; Seleznev 2016).

The aim of this paper is to attract attention to the unjustly underused method of kernel density estimation, to show the author's experience in harnessing the kernel estimator for evaluating the distribution functions in star cluster studies, and to demonstrate its advantages. The paper is organized as follows. Section 2 describes briefly the principle of the kernel estimator. Section 3 summarizes the estimation of the surface and spatial density. Section 4 is devoted to the derivation of luminosity and mass functions, and demonstrates the use of kernel estimator for estimating the velocity distribution and analyzing the Hess diagrams. Conclusions are given in Section 5.

2. KERNEL AND ADAPTIVE KERNEL ESTIMATOR

The principle of the kernel estimator is that every data point is replaced by some function (kernel) normalized to unity. The resulting estimate of the probability density is the sum of all kernels divided by the number of sample points. If the normalization to unity is not needed (that is, if the distribution function is evaluated), it is not necessary to divide by this number. The resulting estimate inherits the properties of the kernel function including its continuity and the differentiability.

The most popular kernel functions are Epanechnikov (quadratic) kernel, quartic (bi-quadratic) kernel and Gaussian kernel (Merritt & Tremblay 1994). These three kernels for one-dimensional case are defined below by Eqs. 1–3 respectively (x in these formulas denotes the distance from the data point). The former two kernels are characterized by finite kernel halfwidth h. Due to this fact the numerical schemes with Epanechnikov and quartic kernels are more efficient (time-saving), because these kernels contribute only to the argument points inside the sphere of radius h. The Gaussian kernel contributes to all points whatever their coordinates, but has higher differentiability. The quartic kernel produces a smoother estimate function than the quadratic one.

$$K(x) = \begin{cases} \frac{3}{4h}\left(1 - \frac{x^2}{h^2}\right), & |x| \le h\,, \\ 0\,, & |x| > h\,; \end{cases} \tag{1}$$

$$K(x) = \begin{cases} \frac{15}{16h}\left(1 - \frac{x^2}{h^2}\right)^2, & |x| \le h\,, \\ 0\,, & |x| > h\,; \end{cases} \tag{2}$$

$$K(x) = \frac{1}{h\sqrt{2\pi}} e^{-\frac{x^2}{2h^2}}. \tag{3}$$

The adaptive kernel estimator uses kernels with a variable halfwidth depending on density. The method consists of two steps. At the first step, the pilot estimate is determined; it is the usual estimate with a fixed value of kernel halfwidth. At the second step, the final estimate is determined. The kernel halfwidth for every data point is calculated as the product of the pilot kernel halfwidth value and the lambda-factor. The lambda-factor is inversely proportional to the square root of the pilot density value for this data point. As a result, the density estimate is derived with the greater kernel halfwidth where density is lower, i.e., in the wings of the distribution function. It leads to a better result for distribution functions with weak wings.

We illustrate this in the left-hand panel of Fig. 1, taken from Seleznev (2016), where the radial profile of the spatial number density of stars is shown for the corona of numerical N-body open star cluster model 1 (Danilov et al. 2014) at the time of about 150 Myr. The solid line is the adaptive kernel estimate, and the open circles show the fixed kernel estimate (it is the pilot estimate for the adaptive method). The adaptive estimate can be seen to be much smoother and unbiased in the area of low density. We set the lambda factors equal to unity at $r < 10$ pc, otherwise the kernel halfwidth of the adaptive estimate became too small and the density profile looked too variable near the center.

The kernel estimator can be easily adapted to multi-dimensional case, see Silverman (1986) and Merritt & Tremblay (1994). However, there is some limitation. The number of points for the appropriate accuracy level of the estimate increases very rapidly with the number of dimensions (Silverman 1986). However, the required number of data points for kernel estimate in 3-dimensional case is still quite reasonable.

A smoothed bootstrap algorithm is used for determining the confidence interval (Merritt & Tremblay 1994). This algorithm involves generating secondary samples distributed in accordance with the distribution function obtained with the kernel estimator and having the same size as the original sample. The secondary samples are modeled using Monte-Carlo (Neumann) method. The original distribution function is approximated by cubic spline, and the differentiability of the kernel estimate is very important here. Fig. 1 (left-hand panel) shows the 2σ wide confidence interval for the spatial radial density profile obtained using 20 secondary samples. The kernel estimate of the distribution function is obtained for every secondary sample. Then the standard deviation is calculated for every point using the density values of the secondary samples.

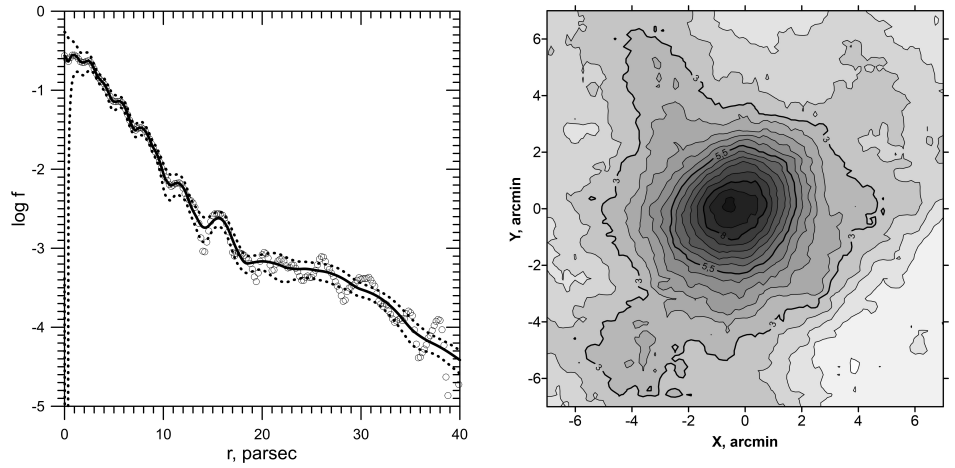

Fig. 1. *Left-hand panel*: the adaptive kernel estimate (the solid line) and the fixed kernel estimate used as a pilot estimate for the adaptive method (the open circles) of the spatial density profile for the corona of N-body model (Danilov et al. 2014). The dotted lines show the confidence interval for the adaptive estimate. *Right-hand panel*: the surface density map of the open star cluster NGC 4337, computed using the quartic kernel with a halfwidth of $h = 3$ arcmin, $V_{\rm lim} = 16$ mag. The density contour lines are marked in units of arcmin^{-2}.

3. ESTIMATION OF THE SURFACE AND SPATIAL DENSITY

Two-dimensional maps of the surface density are very useful tools for the investigation of star clusters. Density values are calculated at nodes of the uniform coordinate grid in the sky plane with the two-dimensional quartic kernel. Then density contour lines (the right-hand panel of Fig. 1) for the projection of the 3-dimensional surface are plotted using some graphics software.

It is important that with the kernel estimation of the density, the latter should not be computed in the h-wide (the kernel halfwidth) strip along the border of the field in order to prevent under-sampling.

The density map provides a general view of the area studied, and it can give a preliminary estimate of the cluster size and structure. It can be used to select control fields (e.g., for deriving the luminosity function of the cluster), to determine the coordinates of the cluster center, to study the cluster structure (e.g., to investigate its ellipticity).

The coordinates of the cluster center are determined as the coordinates of the symmetry center of the innermost density contour (the kernel halfwidth can always be chosen so that the density contour lines corresponding to maximum density would have symmetric shape). The correctness of the center determination can be validated by inspecting the radial density profile. If the center location is determined incorrectly the density profile can have the minimum at the center.

We used the surface density grid to study the distribution of stars of different populations in ω Cen (Pancino et al. 2003). The populations of different metallicity were shown to have different projected distributions. The density contour lines are approximated by ellipses as shown in the left-hand panel of Fig. 2, and the ellipse parameters (axial ratio and orientation of the major axis) are traced

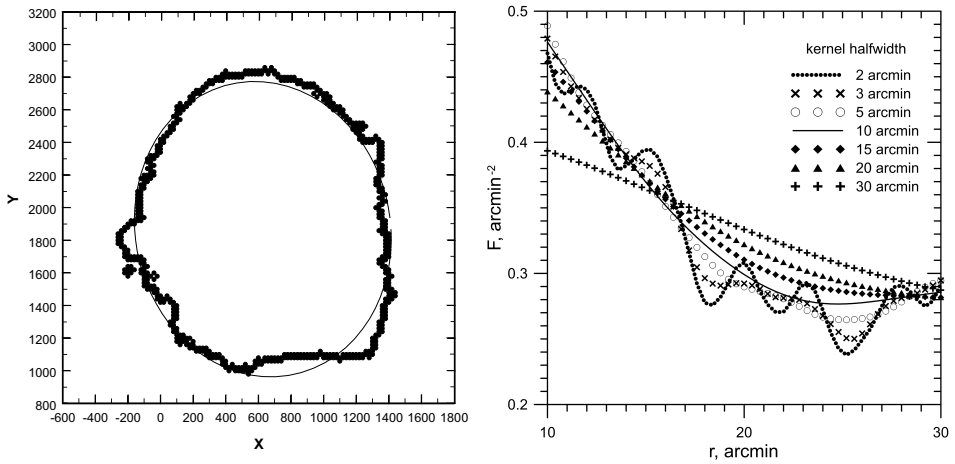

Fig. 2. *Left-hand panel*: approximation of the density contour by an ellipse. The solid circles show the density grid points closest to the selected contour. *Right-hand panel*: selection of the optimum kernel halfwidth for the open cluster NGC 2287, $J_{\rm lim} = 13$ mag.

along the cluster radius.

Merritt & Tremblay (1994) derived formulas for a kernel function for the case of radial surface density profile. They investigated the kernel estimator and maximum penalized likelihood estimator methods in the case of Plummer, de Vaucouleurs, and Michie-King distributions, and showed that the use of the 'optimum' kernel halfwidth determined by minimizing the integrated mean-square error led to unsatisfactory results. Merritt & Tremblay (1994) proposed an empirical approach to the selection of the kernel halfwidths by obtaining a series of profile estimates with different h values and selecting the best one among them.

Seleznev (2016) derived formulas for kernel estimates of the spatial density profile of a star cluster for the case where spatial coordinates of stars (x, y, z) are known. In that study, the spatial density profiles for N-body models of open cluster coronas and surface density profiles for seven open clusters were derived using the kernel estimator.

It is more difficult to find the optimum kernel halfwidth in the case of real star clusters, because the bias is not visible in the outer part of the density profile (every h gives the mean density of field stars, or the background density there). The proposed idea was to use the transition zone between the cluster core and halo (Seleznev 2016). The value of $h = 10$ arcmin was selected in the example shown in the right-hand panel of Fig. 2, because the corresponding density profile is smooth and follows closely the profiles derived with much smaller h.

The surface density profiles can be used to estimate the number of cluster stars and the mass of a cluster (Seleznev 2016). The cluster radius and the mean background density have to be known and a discussion of their estimation can also be found in Seleznev (2016).

Given that the estimated surface density profile is a differentiable function, it is possible to derive the radial profile of spatial star number density. The well-known solution of Abel equation (von Zeipel & Lindgren 1921) can be used to this

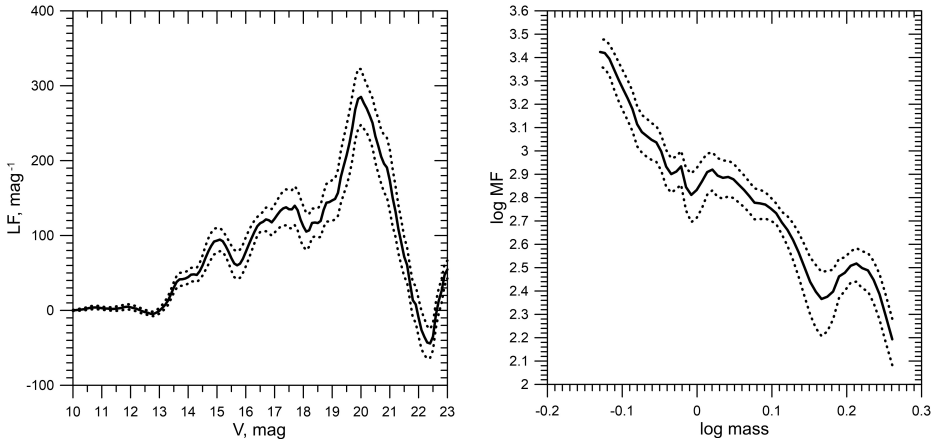

Fig. 3. *Left-hand panel*: the luminosity function of the open cluster NGC 4337. The solid line shows the difference between the kernel estimates for the cluster field and the control field; the dotted lines show the confidence interval. *Right-hand panel*: the mass function of NGC 4337.

end. The spatial density profile can be used, in turn, to estimate the structural parameters of the cluster via Monte-Carlo modeling of its spatial structure. For example, the parameters $\bar{R} =< 1/r_{ij} >^{-1}$ (the mean inverse star-to-star distance), $R_u =< 1/r_i >^{-1}$ (the mean inverse star distance to the cluster center), and $< r^2 >$ (the mean squared clustercentric distance of stars) can be obtained, which are necessary for obtaining the dynamical cluster mass estimates (Seleznev et al. 2016).

4. ESTIMATION OF THE LUMINOSITY AND MASS FUNCTIONS. ANALYSIS OF THE VELOCITY DISTRIBUTION AND HESS DIAGRAMS

We illustrate the kernel estimation of the luminosity function for the open star cluster NGC 4337 (Seleznev et al. 2016). The left-hand panel in Fig. 3 shows the luminosity function of the cluster member stars, obtained as the difference between the kernel estimates for the cluster and control fields. The right-hand panel of Fig. 3 shows the mass function of NGC 4337 obtained from the luminosity function combined with the mass-luminosity relation from the Padova suite of models (Bressan et al. 2012). A linear regression yields a mass function slope of –2.57±0.10.

In the case of NGC 4337 there is a very large discrepancy between the 'photometric' mass estimate based on the mass function and the dynamical mass estimates derived from velocity dispersion (Seleznev et al. 2016). The kernel estimate of the radial velocity distribution function for probable cluster members (Fig.4, the left-hand panel) shows the multi-modal nature of this function. The possible assumption is that only the largest (central) maximum of this function represents the cluster member stars. Other three maxima could be formed by stars of the stellar complex originated in the same star formation act that produced NGC 4337. In this case, all stars selected by radial velocities are located along the same sequence in the color-magnitude diagram (and this, in fact, can be seen in the observed CMD, see Seleznev et al. 2016). If the dispersion of the central Gaussian curve is

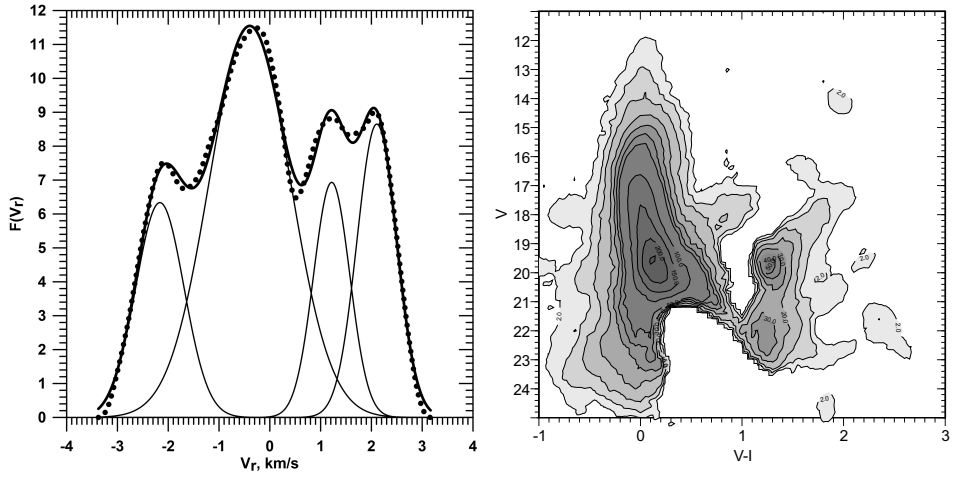

Fig. 4. *Left-hand panel*: the velocity distribution function of NGC 4337. The dotted line shows the kernel estimate and the thick solid line shows the approximation of the distribution function by the sum of four Gaussian curves (the thin solid lines). *Right-hand panel*: the difference between the Hess diagrams of the NGC 2100 cluster field and the control field.

adopted as the (one-dimensional) cluster velocity dispersion then the dynamical cluster mass estimates become much lower, only about twice the photometric mass estimates. Such discrepancy can be explained by the invisible objects – low-mass stars and remnants of the evolution of massive stars.

The kernel estimator can also be used for plotting Hess diagrams – the density distribution maps in the CMD. The use of densities is very helpful in CMDs, because we can subtract the density for the control field from that of the cluster field. The right-hand panel of Fig. 4 shows the result of such subtraction for the star cluster NGC 2100, based on photometric data of Y. Beletsky (2011, private communication). In this case, one can conclude that the control field is not appropriate for the study of the cluster's star content, because it contains many stars located on the main sequence of this cluster.

5. CONCLUSIONS

A kernel estimator can be efficiently used for solving many problems of stellar astronomy in the cases where the distribution function evaluation is needed.

The resulting estimates of distribution functions are continuous and differentiable. It is very important for further application.

A number of codes were written (in Fortran) for different cases of the kernel estimator application. These codes, with detailed instructions, are available from the author upon request.

ACKNOWLEDGMENTS. This work was partially supported by the Ministry of Education and Science of the Russian Federation (state contract No. 3.1781.2014/K, registration number 01201465056). The travel to the conference was supported by Act 211 of the Government of the Russian Federation, agreement No. 02.A03.21.0006.

REFERENCES

Bressan A., Marigo P., Girardi L. et al. 2012, MNRAS, 427, 127
Carraro G., Seleznev A. F. 2012, MNRAS, 419, 3608
Carraro G., Seleznev A. F., Baume G., Turner D. G. 2016, MNRAS, 455, 4031
Danilov V. M., Matkin N. V., Pylskaya O. P. 1985, SvA, 29, 621
Danilov V. M., Putkov S. I., Seleznev A. F. 2014, Astron. Rep., 58, 906
Kholopov P. N. 1953, AZh, 30, 426 (in Russian)
Kirsanova M. S., Sobolev A. M., Thomasson M., Wiebe D. S., Johansson L. E. B., Seleznev A. F. 2008, MNRAS, 388, 729
Merritt D., Tremblay B. 1994, AJ, 108, 514
Pancino E., Seleznev A., Ferraro F. R., Bellazzini M., Piotto G. 2003, MNRAS, 345, 683
Prisinzano L., Carraro G., Piotto G., Seleznev A. F., Stetson P. B., Saviane I. 2001, A&A, 369, 851
Seleznev A. F. 1998, Astron. Rep., 42, 153
Seleznev A. F. 2016, MNRAS, 456, 3757
Seleznev A. F., Carraro G., Piotto G., Rosenberg A. 2000, Astron. Rep., 44, 12
Seleznev A. F., Carraro G., Costa E., Loktin A. V. 2010, New Astron., 15, 61
Seleznev A. F., Carraro G., Capuzzo Dolcetta R., Monaco L., Baume G. 2016, submitted to MNRAS
Silverman B. W. 1986, *Monographs on Statistics and Applied Probability*, Chapman and Hall, London, GB
Vasilevsky A. E. 1985, *Metody Zvezdnoy Statistiki*, UrGU, Sverdlovsk, USSR (in Russian)
von Zeipel H., Lindgren J. 1921, Kgl. Svenska Vet. Akad. Handlingar, 61, 15